AI for Evidential Reasoning

Ludi van Leeuwen[1], Roos Scheffers[2] and Bart Verheij[1]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen. Nijenborgh 9, 9747 AG Groningen

[2]Department of Information and Computing Sciences, Utrecht University. Princetonplein 5, 3584 CC Utrecht

Abstract

We summarize the first workshop of AI for Evidential Reasoning (AI4EVIR), held on December 9, 2025 in Turin, Italy. The workshop was co-located with JURIX 2025, The 38th International Conference on Legal Knowledge and Information Systems.

Keywords

Evidential Reasoning, Bayesian Networks, Belief updating, Hypotheses, Scenarios

Reasoning with evidence to establish relevant facts lies at the heart of legal reasoning. Technologies that allow us to reason about facts are evolving, and new kinds of evidence are becoming available (for example, digital forensic science). At the same time, reasoning with evidence is a dynamic and complex process: practitioners must decide what evidence to collect and how to interpret it amid vast amounts of data. These choices in selecting evidence and the following reasoning with evidence are complex tasks that may benefit from standardization and assistance by AI, for example, to avoid probabilistic and other fallacies.

The AI for Evidential Reasoning (AI4EVIR) workshop aimed to bring together researchers working on evidential reasoning, in the broadest sense, as well as those with expertise in forensic science and evidence evaluation, in order to share their progress on handling various problems. As well as to foster discussion and exchange between theoretical and applied perspectives on how AI can contribute to evidential reasoning in legal and investigative contexts.

We invited submissions of all levels of maturity (early stage, mid stage, completed). A selection was made on the basis of overall quality, relevance, and diversity.

The workshop[1] included an invited talk by Dr. Marouschka Vink, a practitioner at the NFI, who introduced Bayesian networks and Bayesian network methods from forensic practice, and template networks as a method for constructing BNs consistently across experts and cases.

After the invited talk, there were two sessions with presentations of accepted papers in which varied approaches and applications to reasoning with evidence were presented. We started with a presentation by Federico Costantini, presenting on constructing a synthetic dataset, in part generated by LLMs, in order to test the use of LLMs in a digital forensic setting.

AI4EVIR: Workshop on AI for evidential reasoning, December 9, 2025, Turin, Italy.

l.s.van.leeuwen@rug.nl (L. van Leeuwen); r.j.scheffers@uu.nl (R. Scheffers); bart.verheij@rug.nl (B. Verheij)
https://www.ai.rug.nl/~verheij/ (B. Verheij)
0000-0003-3165-4376 (L. van Leeuwen); 0009-0008-1624-3046 (R. Scheffers); 0000-0001-8927-8751 (B. Verheij)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]The CFP, deadlines, PC, and schedule of the workshop can also be found at https://aludi.github.io/AI4EVIR/.

Henry Prakken presented on Bayesian reasoning under rare events, whose presentation focused on how picking reference classes or propositions needs to be done carefully and transparently, in order to avoid (the appearance of) fallacious reasoning.

In the second session, Bertram Ludäscher presented a proposal for a chain of LLMs for trustworthiness, where LLMs together would create an argumentation graph, which could then serve as an artifact for further discussion. The artifact serves as an externally verifiable evidence, increasing the trustworthiness of the model.

Then, Leya Hampson continued with a presentation on the independent creation by two different modelers of a BN model of an entire case, presenting the differences within modelers in going from the qualitative structure of the Bayesian network to quantifying the BN.

Continued, Daira Pinto Prieto discussed advances in a qualitative logic framework for reasoning with evidence, by adding work on the certainty-dominance of evidence. Certainty-dominance is a novel method to compare sets of evidence.

Helen Qiao presented work on comparing human and LLM updating on evidence from the perspectives of the defense and prosecution, and showed that these are conservative Bayesian updaters. LLMs were shown to exhibit a recency bias in various roles and presentation styles.

The talks concluded with a presentation by Henrik Palmer Olsen on evaluating credibility assessments of Danish asylum cases with LLMs, covering the Prompt Valley of Death, an LLM-based categorization of credibility assessments, and an analysis of temporal changes in credibility assessments.

Throughout the talks, main themes that arose were the importance of practice and the extent to which theory should be applied to it; specifying propositions; and whether the study of modeling reasoning with evidence should be a descriptive or prescriptive practice. These themes were further discussed by participants in the wrap-up at the end of the workshop.

The Program Committee (PC) received a total of 9 submissions. Following a single-blind reviewing process, each paper was peer-reviewed by at least two PC members. The committee decided to accept 9 papers, containing original work. One paper was withdrawn due to logistical issues. Two papers are not included in the proceedings on request of the authors. Of one accepted paper, the authors were not able to make it to the workshop. There were 7 presentations and one keynote at the workshop, and there are 6 papers in the proceedings

The specifics of the program were as follows.

Keynote

- Marouschka Vink - *the evaluation of digital findings in forensic casework*

Paper presentations

- Federico Costantini, Fausto Galvan, Francesco Crisci, Luca Baron and Pier Luca Montessoro - *The Quality Assessment of LLM in Digital Forensics*
- Anne Ruth Mackor and Henry Prakken - *On Reporting Likelihood Ratios of Exhaustive and Non-Exhaustive Hypotheses about Rare Events in Criminal Cases*
- Shawn Bowers and Bertram Ludäscher - *Towards Trustworthy AI Results using Evidence Structures: From Certificates to Argumentation Frameworks*

- Leya Hampson and Ludi van Leeuwen - *Investigating the value of qualitative Bayesian networks of complete cases as "double-check" tools on traditional judicial reasoning: An exploratory study*
- Aybüke Özgün and Daira Pinto Prieto - *A Qualitative Logic for Uncertain Evidence and Belief Comparison*
- Mengxuan Helen Qiao, Vanessa Cheung, Leya Hampson and David Lagnado - *Recency Effects, Cautious Convictions, and Conservative Updating in GPT-4o's Legal Decisions (not included in proceedings)*
- Henrik Palmer Olsen, Mohammad N S Jahromi, Frederik Bay-Jørgensen, Thomas B Moeslund and Thomas Gammeltoft-Hansen - *Managing Fuzziness: Leveraging LLMs for Discovering Credibility Indicators in Asylum Cases (not included in proceedings)*
- Mario Guenther and Conrad Friedrich - *Probabilifying the Scenario Approach to Legal Proof (Unable to present)*

Organization

Workshop Chairs

- Ludi van Leeuwen, University of Groningen
- Roos Scheffers, Utrecht University
- Bart Verheij, University of Groningen

Program Comittee

- Floris Bex, Utrecht University
- Christiaan Dahlman, Lund University
- Marcello Di Bello, Arizona State University
- Hylke Jellema, Utrecht University, University of Groningen
- Jeroen Keppens, King's College London
- Anne Ruth Mackor, University of Groningen
- Daphne Odekerken, Netherlands Police
- Annet Onnes, Utrecht University
- Henry Prakken, Utrecht University
- Silja Renooij, Utrecht University
- Burkhard Schafer, University of Edinburgh
- Cor Steging, University of Groningen
- Rineke Verbrugge, University of Groningen
- Marouschka Vink, Netherland Forensic Institute
- Amy Wilson, University of Edinburgh

Acknowledgments

The organisation of the workshop was supported by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. The organisation of the workshop was also supported by the project 'AI4Intelligence: From Multimodal Data to Trustworthy Evidence in Court' with file number KICH1.VE01.20.011 of the research programme Data and Intelligence, which is partly financed by the Dutch Research Council (NWO). The organizers would like to thank the JURIX 2025 workshop chairs and organizers for providing an excellent framework for our workshop.